DOCUMENT RESUME

ED 155 638 CS 004 156

AUTHOR Calfee, Robert C.; Drum, Priscilla A.
TITLE How the Researcher Can Help the Reading Teacher With Classroom Assessment.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 76
CONTRACT 400-75-0049
NOTE 68p.; Paper presented at the Conference on Theory and Practice of Beginning Reading Instruction, Univ. of Pittsburgh, Learning Research and Development Center, May 1976; For related documents see, CS 004 132-133, CS 004 135, CS 004 137-173, ED 125 315 and ED 145 399, Some pages may reproduce poorly due to small type

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Beginning Reading; Conference Reports; Decision Making; Elementary Education; Primary Education; Reading Comprehension; *Reading Instruction; *Reading Research; *Reading Tests; *Teacher Responsibility; *Test Construction; Testing; *Test Interpretation; Tests

ABSTRACT

This paper discusses some specific issues about testing and relates the discussion to reading teachers and reading instruction. The issues that are discussed include the goals, criteria, and methods of classroom assessment. The paper concludes that teachers need to learn more about the process of assessment in order to use tests effectively, and that classroom assessment ought to aim toward the precise and efficient measurement of specific component skills for short-term decisions. Discussion following presentation of the paper is included. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED155638

How the Researcher Can Help the Reading Teacher With Classroom Assessment

BEST COPY AVAILABLE

Robert C. Calfee

Stanford University

School of Education

Priscilla A. Drum

University of California, Santa Barbara

This paper was presented at the conference on Theory and Practice of Beginning Reading Instruction, University of Pittsburgh, Learning Research and Development Center, May 1976.

CS004156

Conferences supported by a grant to the Learning Research and Development Center from the National Institute of Education (NIE), United States Department of Health, Education, and Welfare, as part of NIE's Compensatory Education Study. The opinions expressed do not necessarily reflect the position or policy of NIE, and no official endorsement should be inferred. NIE Contract #400-75-0049

ACKNOWLEDGMENTS

Preparation of this paper was supported in part by a grant from the Carnegie Corporation. Many of the ideas are reported in a slightly less technical form in "What Research Can Tell the Reading Teacher about Assessment," by R. C. Calfee, P. A. Drum, and R. D. Arnold, to appear in S. J. Samuels (Ed.), What Research Has to Say to the Teacher of Reading, an IRA publication. The assistance of Elizabeth Orem, Dorothy Piontkowski, Barbara Tanner, and Barbara Tingey is gratefully acknowledged. A special debt is owed Kathryn Hoover Calfee for her practical counsel about the classroom.

How the Researcher Can Help the Reading Teacher with Classroom Assessment

Robert C. Calfee
Stanford University

Priscilla A. Drum
University of California, Santa Barbara

To many educators, tests seem an unavoidable nuisance. Though they are useful to some people for certain purposes, increasingly their usefulness and appropriateness are questioned. A rising chorus questions whether tests really provide fair and useful measures of educational progress, and colleagues caution against overuse of tests to no good purpose (e.g., Venezky, 1974a; Levine, 1976).

The measurement tradition is strong in educational psychology. Tests are one of the few "scientific" elements in educational research and practice, and they can serve a vital role--evidence is essential to effective and efficient instruction. For instance, there are definite limits to what a lecturer can hope to achieve, because he obtains relatively little information from the members of his audience--he must rely on eye contact, on signs of attentiveness, and on questions from the listeners. At the other end of the continuum, individualized instruction builds on the continual exchange of information between teacher and student; the instructional program is continuously realigned to the student's needs and strengths (e.g., Atkinson & Paulson, 1972). Frequent, precise, and appropriate assessment is critical to this process. But such testing must be designed to fit the instructional needs of the teacher--this is the burden of our present message.

The testing tradition, following the lead of Alfred Binet, has focused attention on the selection and sorting of individuals. One can find occasional

comment on "teacher-made" tests in books on educational testing. But even here the criteria are those applied to tests for selection and sorting, by and large. That other needs exist is reflected in the plethora of terms--criterion-referenced, domain-referenced, behavioral-objectives, diagnostic--all denoting something other than the conventional testing approach.[1] There is little evidence that the "new" tests do a different job from the old ones; it is also worth noting that tests with quite different labels look quite similar. Researchers can provide a service to teachers by looking systematically at the needs for assessment in the classroom, and by analysis of the theoretical and empirical issues in this area. The goal of the present paper is to suggest to researchers some specific issues that warrant investigation; a companion paper has been prepared to look at these issues from the teacher's perspective (Calfee, Drum & Arnold, in press).

An Overview and Two Conclusions

How can assessment be tailored to fit the needs of the classroom? To answer this question, we need to consider three other questions:

(1) Assessment for what? (The goals)

(2) How to assess? (The methods)

(3) Is assessment doing its job? (The criteria)

We will consider in turn each question as it relates to reading, but first, two major conclusions:

(1) Teachers need to learn more about the process of assessment in order to assess for instructional purposes.

(2) Classroom assessment ought to aim toward the precise and efficient measurement of specific component skills for short-term decisions.

The bulk of the paper will buttress and illustrate these two generalizations, but some preliminary comments will help to set the stage. Much research

on reading assessment aims toward goals quite different from those that are foremost for the classroom teacher. The goal of conventional achievement test construction (and thus of the research that centers around such test construction) is the measurement of reliable and substantial individual differences, based on stable scores for each student which place him at some point below, at, or above the average for some larger population. The aim is an instrument for making major, long-term decisions about students, teachers, and programs (Carver, 1974).

The teacher needs information of a much more immediate character. How well can the student read now? What specific reading instructions should he receive next? Is the instruction successful? To the question, "What does the research literature on assessment have to say to the classroom teacher about his instructional needs?", the answer is "Not much!"

To our knowledge, no existing assessment system handles the range of assessment tasks encountered by the reading teacher. Most commercial tests provide little evidence useful for instruction, and are too expensive in time and effort for the teacher's needs. It is not that commercial tests are faulty, rather that they are designed for other purposes than immediate instructional decisions.

Moreover, the researcher cannot focus attention solely on the characteristics of the assessment system, if his goal is the proper assessment of students in the multifaceted happenings of the classroom. The researcher must plan investigations where variation in the characteristics of the assessment system are only one set of factors--the design must also call for variation in the teacher's background and training, in the makeup of the class, and in the nature of the instructional program. Designs of this comprehensiveness require more thoughtfulness than has been typical of educational research, but they are technically feasible (Calfee, 1975a).

In fact, one can argue that research on classroom assessment should not center on test construction at all, but rather on teacher training. Public and private contractors will hopefully improve the kinds of assessment systems available to the teacher. But we suspect that the key to adequate assessment for instructional decision-making in the classroom is the classroom teacher who knows how to select with care from what is available for other purposes, who can modify and simplify the materials at his disposal with an eye to practical application. If so, the chief task of those who would improve assessment of reading for purposes of instruction lies not in psychometrics, but in improving teaching. This does not mean that all psychometric problems have been solved, to the contrary. It simply means that psychometrics may not be at stage center.

Goals of Assessment

What are alternative goals in assessment? First, certain goals aim toward long-term prediction. This is true in evaluation of the individual (Cronbach & Gleser, 1965), for job placement, for school admission, for a grade or achievement mark of some sort. It is true when assessment serves for evaluation of a program. The administrator has to decide whether a curriculum is effective, whether a special program is better than the regular program, whether extra money is making a difference. Diagnosis also falls in this category. Diagnosis is for special cases, like physical anomalies. A person who can't see well has trouble learning to read. If he can't hear very well, he may also have trouble in school tasks, including reading. These are special cases and may require a clinical specialist.

Other goals aim toward short-term decisions. Assessment can serve for instructional decision-making by the classroom teacher. The instructor has to stay current on what each student knows if instruction is to be precisely

directed toward specific needs. "Individualization" is the usual label for this concept. Each student is assessed as to his present skills, abilities, and knowledge, so he can be helped to move from where he is toward some reasonable goal.

Tasks other than individualization also require the classroom teacher to apply skills in assessment:

> It is the beginning of the school year. The teacher is new to the school, and wants to supplement information in the "cum" folder with his own evidence.
>
> A new student arrives in class at midyear, and there is little information available on how well he can read.
>
> The teacher is planning to introduce a new topic (e.g., how to handle polysyllabic words), and needs to know which students know something about the topic, and which ones are totally unprepared.

In summary, assessment for short-term instructional decisions covers diverse situations: (a) optimizing instructional sequences, (b) measuring immediate response to instruction, (c) regrouping for instruction for specific purposes, and (d) deciding on selection and allocation of resources (who needs the aide's time, the tutor's time, the terminal's time?).

Present Methods of Assessment in Education

Psychometrically "sound" tests in use today include normative and criterion-referenced tests; these two types of tests differ little in content, though designed for different applications (Green, 1976). A norm-referenced test shows how the student's score or the class's score compares with the other students or classes who provided the standards for the test. A criterion-referenced test provides a score for a student or a class based

on the number of items mastered (answered correctly) compared with some absolute standard. Neither type of score tells the teacher what a student knows or does not know; direction for further instruction is not indicated. Both types of tests are standardized; an exact procedure for administration is called for, with little room for clinical probing. Most tests are group-administered and use a multiple-choice format to facilitate machine scoring. The content resembles "goulash;" though a subtest structure is often imposed on the test items, the high intertest correlations belie the different names assigned to subtests.

It can also be said of these tests that they are reliable, that the student's relative standing is stable over time, and that they are highly predictive of one another (Bloom, 1964). They are time consuming to administer; they are generally not capable of repeated administration--two or three times a year at most. They yield a single type of measure (percentage correct or some transformation thereof).

Such tests have been developed to meet certain implicit and explicit criteria. It therefore makes sense to consider the standards and criteria that apply to the construction, administration, and interpretation of a test.

Criteria for Evaluating Tests

We want to examine briefly several criteria for evaluating tests: reliability, validity, appropriateness, independence, discriminability, cost, and repeatability. The first two are usually discussed in texts on testing, the others generally not (e.g., Anastasi, 1968; Cronbach, 1970; Farr & Tuinman, 1972). Each criterion has several facets to it.

Reliability: Does the Instrument Provide a Consistent Measure?

In general, reliability refers to the degree to which a measurement is consistent. We can consider the consistency in performance when a person is tested

with one form of a test and then retested with a slightly varied form. Several things have changed. The exact form and content of the test have changed. The student has probably changed. He may have learned something, he may have forgotten something, he may have a headache now that he didn't have earlier. All these sources of variability tend to reduce the reliability in test-retest situations.

Test developers tend to emphasize within-test reliability. There are a variety of ways of thinking about this form of consistency (Cronbach, 1970, Ch. 6). For instance, suppose you divide the items at random in two and correlate the two subscores. Repeat this operation for all possible split-half divisions of the test, then compute the average correlation between the half-scores (Cronbach, 1951). This provides a measure of the extent to which each item contributes consistently to the total test score. One way to obtain "perfect" intratest reliability is to use a test in which the student either fails all items or passes all items. Test developers, to the degree that they strive for intratest reliability, are under pressure to eliminate test items that yield divergent patterns of performance from one student to the next. The items that remain seem likely to measure general performance characteristics rather than performances that reflect specific intructional outcomes. So if you want a perfectly reliable test, ask the same question twenty times. Either a student knows the answer or he doesn't. This would be absurd, of course, but in the limit it is the "ideal" toward which reliability aims.

Maximizing intratest reliability is important when the test score is to serve for a major decision, but it may be counterproductive for instructional decision-making. Teachers need to know more than the student's general ability. Individualization requires knowledge of diverse patterns of performance on specific tasks for different students. For the teacher, a "reliable" assess-

ment instrument is more properly defined as one which accurately and consistently indicates the specific patterns of instruction that best fit the student's needs and capabilities. We shall examine this matter in more detail later in this paper.

Validity: Does the Instrument Measure What It's Supposed to?

As with reliability, the concept of validity assumes many guises. Face validity means that the test looks like it measures what ought to be tested. Construct validity means that if several tests seem to be measuring the same thing, there must be something there to be measured. Predictive validity means that there is a correlation between a test and a criterion of performance (usually another test).

To possess adequate validity for most educational purposes, a test usually has to satisfy each of these criteria. For instance, one can predict reading achievement reasonably well from mathematics tests, but teachers and parents would question the face validity--it would not be seemly to measure reading performance with a test containing arithmetic "word problems," even if the test met the usual standards of reliability and predictive validity. The researcher could provide a service by exploring the issue of instructional validity--a test is valid when it points to an instructional treatment that improves the student's performance on a specified task. From this point of view, aptitude-treatment interaction research aims to validate various aptitude tests (Cronbach & Snow, in press; Walker & Schaffarzick, 1974).

Appropriateness: Does the Instrument Measure Sensibly, Given the Use to Which the Evidence Is to Be Put?

Appropriateness is introduced here as a fuzzy concept covering several related matters. In part, it has to do with whether a test is linked to the goal of an instructional program with sufficient directness and breadth. Researchers

learn the meaning of this concept when public school teachers ask why no one tests what they teach. This complaint is fair and deserves the attention of evaluators.

Appropriateness is disregarded in the common practice of assigning a student to a particular level of a test according to his age or nominal grade placement rather than his actual performance level. The experience of the Chicago schools when they selected achievement tests to be appropriate to the students' actual reading level is instructive in this connection (Chicago Tribune, 1975). Asking a high school student who reads at the first or second grade to handle an advanced level of the Metropolitan Achievement Test is a mistake; whatever the score, it is unlikely to reveal the student's actual skill in reading. The advanced test is for those reading at grade six or above. Students reading at a lower level are likely to guess randomly at the answers, but this performance is likely to lead to a grade level score that is higher than their actual reading ability.

Finally, appropriateness seems to distinguish many conventional academic achievement tests from the alternatives represented in the National Assessment of Educational Progress. The goal of NAEP was to cover the range of reading tasks that a literate person might confront in his experiences in school, at work, at play, and in the other aspects of life in the society. The typical comprehension test is simply not appropriate to cover the broad array of "themes" that seemed important to the NAEP staff:

1. understanding words and word relationships (literal comprehension of isolated words, phrases, and sentences);
2. graphic materials (comprehension of the linguistic components of drawings, signs, labels, charts, maps, graphs, and forms);

3. written directions (comprehension of directions, plus ability to carry them out operationally);
4. reference materials (comprehension and knowledge of indices, dictionaries, alphabetizing, and TV listing formats);
5. gleaning significant facts from passages (comprehension, and to a limited extent, recall, of literal content in the context of a larger reading passage);
6. main ideas and organization (ability to abstract upwards from the sentence-by-sentence content of a passage and recognize main ideas and organizational features);
7. drawing inferences (ability to reach a conclusion not explicitly stated in the passage, in most instances relying only on information given but in a few cases on knowledge unrelated to the passage);
8. critical reading (ability to recognize author's purpose, and to understand figurative language and literary devices) (Mellon, 1975).

It is also the point of the research of Sticht and his colleagues (Sticht, 1975; Sticht, Caylor, Kern, & Fox, 1971) that the assessment of a person's reading ability (and the preparation of what he is expected to read) should be appropriate to the task demands--don't make life unnecessarily difficult by asking hard, tricky questions when easy, plain ones will do.

Independence: If Several Skills Are Measured, Is There Evidence That They Are More or Less Separable and Autonomous--Not Closely Correlated?

To be most useful, the several scores from an assessment battery should provide the teacher with distinctive pieces of information. When all the subtest scores are highly intercorrelated, the teacher receives little guidance about distinctive courses of action. As Thorndike (1973) has pointed out, even a modest degree of correlation between two scores (r = .6 or more) makes

it difficult to make differential diagnosis, given that the scores are normally distributed. The magnitude of this problem for certain commercial tests has been discussed by Calfee and Venezky (1969), and possible remedies suggested (Calfee, in press). One desirable condition is that each test be "clean", i.e., that steps be taken to insure that the test measures the desired skill and none other. We will describe later a second approach built upon factorial test design, in which systematic variation in the materials and conditions of testing allows the tester to find out the circumstances under which a student can and cannot handle a task.[2]

Discriminability: When Possible, Information from a Subtest Should Be "Yes-No."

It takes more expertise and attention to monitor an ammeter and make decisions about an automobile's electric system than to notice simply whether the generator light is on or off. Similarly with a test--when the scores on a test take the form of a normal distribution, then fine gradations in performance matter a lot and interpretation is more difficult. It is much easier to interpret performance when it is either clearly at the mastery level or altogether faulty, with no "in between" scores. Careful specification of the task is required, but the benefits for instructional decision-making can be considerable (Calfee, in press).

Cost: How Much Time and Money to Buy, Administer, Score, and Interpret?

Tests cost money, and they cost time. These costs may be overlooked by teachers, even when they are the ones who pay. For instance, in one school, teachers spent three days testing the students' reading skills in third and fourth grades. The scores were then used for the sole purpose of sorting students into three reading groups: high, medium, and low. Obtaining a ten-minute oral reading sample from each student would probably have done the sorting job as well, or better, and at much less cost. When a major decision is to

be made, substantial cost is justified; when continuous short-term decisions are required, low cost is essential.

Repeatability: For Classroom Instruction--It Should Be Practical to Readminister a Test Whenever the Teacher Needs Information.

The time and cost required by many tests makes repeated administration impractical. Besides this, the psychometric concern with reactivity in retesting leads to advice against repeated administrations of the same form. It is rare to find more than two alternate forms of most commercial tests. For evaluation of a program or an individual, assessment once or twice a year is sufficient. But the teacher who wants evidence on the effectiveness of yesterday's instruction needs an "off-the-shelf" test, one which comes in many forms, and can be used as often as necessary.

A Closer Look at Reliability

If any concept is central to research on assessment, reliability certainly seems the candidate. As noted above, in its simplest form reliability means that a measure is consistent and reproducible. Suppose, when a carpenter used his ruler to measure the length of a board, that each "inch" on the ruler acted somewhat differently during the measurement process. Then the results of the measurement would vary depending on which particular ruler was used and the length of what was being measured, among other things. This is manifestly undesirable. By analogy, the designer of a test for the measurement of academic outcomes seeks to build a test from a set of items that act together consistently to measure the skill or knowledge of interest. Indices of intratest reliability such as split-half reliability, the point-biseral coefficient, alpha, or the KR-20 index reveal the extent to which performance on each item in a test contributes in a consistent fashion to the total score.

Another way of thinking about reliability builds on the analysis of variance procedure (Cronbach, 1970, pp. 158ff). For instance, consider the scores for five students in Table 1. These records show a fair amount of consistency. Students may do well or poorly, but each item contributes consistently to the total score. Item 4 is harder than the other items, and the students who do most poorly always do poorly on this item. Similarly, Item 1 is relatively easy, and consistently so for the students who do best.

Insert Table 1 About Here

The magnitude of the consistency can be determined through the standard analysis of variance (refer to Cronbach, 1970, p. 159, for details of the procedure). The total variance in the scores can be partitioned to yield three variance estimates (Table 2). The expected value of each variance estimate allows one to compute the variance component for each source, as shown beneath the analysis of variance summary table. Thus, the variance of the students' "true" scores is estimated to be $\sigma_S^2 = .487$; the variance in the student-item interactions, σ_{SI}^2, is estimated to be .113. The student total-score variance is a measure of individual differences in the total scores. The student-item interaction is a measure of inconsistencies in the way different students react to different items. In this example, the idiosyncratic variation in items is relatively slight, compared with total score variance. As an index of the consistency of the contribution of individual items to the total score, Cronbach (1951) proposed the ratio of true score to observed score variances. This is equivalent to the ratio between total score variance and overall variance (total score variance plus idiosyncratic student-item variance):

$$\alpha = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_{SI}^2}$$

The principle here is quite simple--to take seriously the student's total score as an index of individual difference, variation in the set of "true" scores should account for a fairly large proportion of the overall variance in the observed scores, which can be shown to be the sum of σ^2_S and σ^2_{SI}. As can be seen at the bottom of the table, the Cronbach alpha for these data, $\alpha = .81$, is quite high, given the limited data.

Insert Table 2 About Here

Incidentally, what is estimated as a reliability in this example, and throughout the discussion that follows, is what Cronbach (1951) calls α_1, the consistency of the contribution of the individual item to a summary score of some kind. One can also calculate the overall reliability of the total score of a test or subtest, but for our present purposes it is item reliability that is most important. It should also be mentioned that the estimates of α in this discussion are biased; the procedure for calculating unbiased estimates is straightforward (Winer, 1971, p. 282), but would unnecessarily complicate the example. Finally, no effort is made to apply the Spearman-Brown correction for test length.

As an example of an inconsistent set of items, consider the student-item matrix in Table 3. The variation in the total scores of individual students is exactly as in Table 1, but if you examine the data closely, you will see that the items are less consistent. Items 1 and 2 are passed by some of the students whose total score shows many errors; the same items are failed by some of the students whose total score shows many successes. These idiosyncratic reactions of particular students to particular items in an unpredictable and inconsistent manner are referred to as subject-item interactions. The estimate of student-item variance is indeed higher for this matrix (MS(SI) = .200), and the reliability is .67, or 20 percent less than the results in Table 2.

Insert Table 3 About Here

What are the characteristics of a test with a high reliability coefficient? First, there must be individual differences of substantial magnitude in overall performance. This is another way of saying that σ_S^2 must be relatively large. Second, idiosyncratic reactions to particular items by individual students must be small; put otherwise, σ_{SI}^2 must be relatively small. Items that do not fall into line are relatively easy to detect, and the dependability of the student's total score is markedly improved by eliminating those items that do not fall into line. For instance, if Items 1 and 2 are eliminated from the test in Table 3, the test becomes perfectly reliable.

Suppose, however, that the purpose of the test is not to generate a single total score, but to yield patterns of performance, which might serve usefully for specific instructional responses. We will show now that the conventional approach emphasizing total-score reliability can lead to the elimination of the items that provide the essential information about such patterns. However, extensions of the same basic procedure for determining reliability can be used to evaluate the dependability of those patterns that do exist in the data. These extensions build upon the landmark work of Cronbach and his colleagues on generalizability theory for psychological assessment (Cronbach, 1951; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; for a different perspective on a similar problem, see Calfee, 1976; Calfee & Elman, in press).

The key to the evaluation of patterns of individual differences is to think about the reliability of the patterns, rather than the reliability of the total test score or of a particular subtest score. The analysis of variance technique provides the technology to support this thinking, which is why we introduced it earlier. The concepts will be introduced with the aid of a

specific example, the student-item matrix in Table 4 (disregard the subtests for now). Suppose a teacher has developed an eighty-item test, and has collected scores from twenty students. He gives you, the researcher, the data and asks that the reliability of the test be determined, to insure that the instrument meets customary standards. We will now proceed to examine these data in some detail. At first glance the test will appear relatively unreliable. However, closer attention to the structure of the data--a process much like peeling an onion--will uncover a great deal of reliable information. The analysis of variance will provide a systematic accounting of the information, and at each stage of analysis we will see that reliability coefficients of increasing specificity will be determined.

Insert Table 4 About Here

Casual examination of the matrix in Table 4 shows you that, while there are substantial individual differences in the total scores, there is also considerable idiosyncratic variation in the reaction of particular students to particular items. The situation is not too bad, as can be seen from the analysis in Table 5. The reliability measured by α is of a respectable magnitude by many standards, especially when you remember that the α value in Table 5 is the reliability of a single item. The α value for the total score can be shown to be $\alpha_k = .93$, for instance (Winer, 1971, pp. 286-287).

Insert Table 5 About Here

The test designer then remarks to you that the test actually comprises items from two distinctive categories, and he is curious about whether the two subtests reveal the differences in performance they were designed to measure. In Table 5 the items can be arranged according to a subtest structure. If you look at the first four and last four items for each student, you can see more consistent patterns within each subtest than appear when the test is examined as a whole. Each student tends to succeed or to fail on all the items within a subtest—there is only modest deviation from the all-or-none pattern. This

suggests that the performance patterns are more reliable than the overall measure in Table 5 suggests.

In Table 6 is shown the determination of reliabilities for this situation. The analysis of variance now includes the Test factor as a source of variance, along with the Student-Test interaction. Two reliability indices can be computed, in answer to the questions:

- Oo How consistent is the contribution of the subtest score to the total score? The answer is, only slightly so, $\alpha = .162$. (The subtest is the "item" in this analysis.) Look at the data and you will see that some students have a high score on T_1, some a low score; some have a high score on T_2, some a low score; and all combinations of high and low on each subtest are represented. In other words, there are substantial student-subtest interactions.
- Oo How consistent is the contribution of each item within a subtest to the difference between the student's subtest scores? The consistency here shows up in this reliability coefficient, $\alpha = .836$.

Insert Table 6 About Here

The increase in the last mentioned reliability coefficient compared with the total-test coefficient in Table 5 seems modest; only about 10 percent. But there is a substantial gain in our understanding of the test structure--we can see that individuals differ considerably in the subtest patterns, whereas the total test score is not reliable compared with variations in subtest--student interactions.

What does the preceding analysis of reliabilities tell the test designer in this particular instance? The overall reliability of the test (Table 5) is moderate, but not spectacular. From this analysis alone, the test designer might be advised to throw away some of the items that contribute least consistently to the overall test score. This would be a mistake, because these same

items contribute most consistently to the subtest patterns. The subtest analysis (Table 6) reveals that the subtests themselves contribute inconsistently to the total test score, but the items within each subtest yield fairly consistent patterns of individual differences in the subtest scores. These patterns are readily visible to the naked eye. To be sure, we created the data set, and so we knew what the underlying structure really was. But there is an important moral: it behooves the test designer to think seriously about the dimensions of the test, and of the characteristics of the students for whom the test is being designed (Calfee, 1976).

We have illustrated how the researcher can help the teacher in the conduct of classroom assessment, using one of the oldest tools of the educational psychologist's trade--the analysis of test reliability. To be sure, more is needed than the examination of reliability of a total score. The tools exist today for the investigation of the reliability of structural patterns, and it is these that are likely to be of service to the classroom teacher. Incidentally, the payoff from structural analysis increases with the complexity of the structure. The test in the example above had the simplest possible structure--two subtests. As the number of independent dimensions of pattern increases, and as the number of student groups for which these are useful dimensions increases, it becomes more important that the researcher turn away from simple "omnibus" reliability to the more precise investigation of structural reliabilities.

The Instructional Validity of Simple Decisions

After reliability, the second cornerstone of test theory is validity. We want to consider here some ideas about the validity of decisions based on test results, where a major consideration is the simplicity of the decision. A decision in this context is a prediction--based on the evidence, the student

is likely to succeed if the situation remains as is, or the student is likely to fail unless something out of the ordinary is attempted. One could also enquire whether the test points with accuracy toward a specific instructional treatment, but we will not deal with that issue here.

The usual approach to prediction in educational settings is the venerable Pearson correlation coefficient. It assumes that two normally distributed covariates share some common variance in the form of a linear relation. This solution is elegant and most teachers learn something about correlation during their preservice training.

The technique is straightforward. If we know (a) a student's score on the predictor test, A, (b) the mean and variance of A, (c) the correlation between A and the criterion or to-be-predicted test, B, and (d) the mean and variance of B, then we can readily compute an estimate of the student's probable performance on B, along with confidence bounds on the estimate. This procedure assumes normality of the distribution of scores.

Teachers seldom make use of the procedure just described. They are not comfortable with statistics, they have neither the time, the information, nor the computational formula. Thus, knowing that .70 is the correlation between a child's score on a readiness test at the beginning of kindergarten and his first-grade reading achievement is little help to the typical classroom teacher, no matter how dedicated he might be. Of even less help are predictive relations established by more sophisticated techniques, such as step-wise multiple regression, discriminant analysis, factor analysis, or the like.

In our research we have explored some alternative approaches to prediction based on all-or-none tests, with interesting consequences.[3] The general technique is most conveniently presented by a concrete example. A kindergartner's knowledge of the names of the letters of the alphabet is known to be predictive

of subsequent performance on reading achievement tests. (See for example, Gibson & Levin, 1975; Venezky, 1975). The reasons for this relation are complex, and undoubtedly have more to do with home environment, general ability, amount of time spent watching Sesame Street, and so on, than with specific training on letter names. Alphabet knowledge is an indicator, not a cause, of reading success or failure.

The technique works as follows: early in the school year ask a group of kindergartners to name each letter of the alphabet--this yields the predictor score. What shall we predict? Suppose we measure reading achievement of these children two years later when they leave the first grade. Divide the students into two groups: those who read at or above grade level and those who are below grade level. The former group has "succeeded" by conventional standards. The children in the latter group are below an acceptable level of performance, and might have profited from additional instruction during kindergarten and first grade. In any event, we have a simple metric to be predicted--success or failure.

Now for the validation. How well can the kindergarten teacher sort children into those who will probably succeed and those who probably need additional help, using the child's knowledge of letter names? What is the decision rule for sorting; how complicated does it have to be; how accurate will it be?

We have some data on this question. Kindergarten children were tested in 1970 on their ability to name each of the twenty-six upper-case English letters (Calfee, in press). Two years later at the end of first grade, they took the Cooperative Primary Reading Test (Educational Testing Service, 1970). We obtained complete records for 144 children from the original sample of 276. There is a marked relation between alphabet knowledge and reading achievement in this group of students; the correlation is .50.

An interesting pattern appears if we examine the frequency distribution of alphabet scores (Figure 1). First, the distribution for the entire sample is markedly bimodal (top panel). Second, children who are below grade level at the end of the first grade are disproportionately represented at the lower end of the distribution (they did not know their ABCs at the beginning of kindergarten), whereas the children who were above grade level are disproportionately represented at the upper end (they did know their ABCs). The correlation describes accurately the linear relation between the two variables, but it does not reveal the bimodality of the distributions and the potential for simple decision-making inherent in that bimodality.

Insert Figure 1 About Here

In particular, suppose we sort children into two groups by a "cut-point" on the alphabet knowledge distribution; we might classify as "in need of additional instruction" all children who identified ten or fewer letters. Then 12 of the 61 children who were at or above grade level would have been misclassified as needing additional instruction (they knew ten or fewer letters when they entered kindergarten, but met the grade level criterion at the end of the first grade); 28 of the 84 children who were below grade level would have been misclassified as <u>not</u> needing additional instruction (they knew more than ten letters on entry to first grade, but failed to meet the grade level criterion). This means that by placing a cut-point at ten or fewer letters correctly identified, 12 out of the total 144 students, or 8 percent, would be misclassified as needing instruction when they would end up doing all right without it, and 28 out of 144, or 19 percent, would be misclassified as not needing instruction, but would end up below criterion. The total misclassification rate would thus be 27 percent at this cut-point.

Figure 2 shows what happens as the cut-point is moved from the lowest to the highest alphabet score for this set of data. If the cut-point is at the

extreme left of the abscissa, then even if a child cannot identify a single letter, he is given no supplementary instruction. All of the children who fail to reach criterion are misclassified under this condition; none of the children who meet criterion are misclassified, of course, since by definition they need no additional help. As the cut-point is moved to the right, more and more students are assigned to supplementary instruction. At first, most are from the below-criterion subgroup. There is a wide flat spot in the misclassification function, reflecting the small number of students in the middle portion of the bimodal distribution of alphabet knowledge scores. At a cut-point (or critical value) of 10 in the figure, the percentages mentioned above can be seen; 8 percent of the students are falsely classified as needing more help, 19 percent of those that need help are not so classified, for a cumulative misclassification rate of 27 percent (the sum of the previous two percentages). Eventually, at the right-most side of the abscissa, all students receive supplementary instruction, even those who know all the letter names. This means that all of the above-criterion students are, by definition, misclassified.

Insert Figure 2 About Here

Let us emphasize two features of this procedure. First, it is simple. We can say to the teacher: "Give the child a test. If he makes more than X successes, he's probably (this can be made more precise) going to do all right. If he makes X successes or less, then he's probably going to be in trouble and you had better think about what might be done to prevent failure." There are no complicated statistics.

Second, it is robustly accurate. The total misclassification rate in Figure 2 drops to a low of 25 percent, and stays at that level over a broad range of cut-points. (Incidentally, Feshbach, Adelman, & Fuller, 1973, using a predictive test battery, or teacher judgment, or both, found that the misclassification rate from their measures and procedures ranged around 25 percent for a sample of almost 600 students.)

It should be stressed that nothing in the present analysis of alphabet knowledge scores and reading achievement is implied as to the most appropriate action for a child in need. This is clearly not a precise test that calls for a specific treatment. It is probably acting as a general indicator of a variety of abilities and skills; the instructional response can be only a general one.

Standards for Practical Classroom Assessment

A cursory examination of the research literature reveals the emphasis on tests suitable to long-term, major decisions (e.g., Weintraub et al., 1974, pp. 460-464; 1973 pp. 429-447). The teacher's need for in-class assessment, on the other hand, is best met by tests that are speedy, precise, clearly "appropriate," and flexibly repeatable. The concepts of reliability and validity need to be defined in unconventional ways to serve in the design of tests for instructional decision-making.

The teacher cannot expect to find on-the-shelf tests that are well suited to short-term instructional decisions. Moreover, training on "test construction" reflects the conventional psychometric tradition, and so the teacher is likely to be poorly prepared to select, to adapt, and to create useful instruments. It is not the intention of this paper to go into detail about the program of teacher training that might alleviate this gap. However, we suspect that it would center about an analytic approach to "what is being taught"--we have referred elsewhere to the distinction between a "jello" model of the mind in contrast with the "works in a drawer" model, the former being more Gestalt-like, the latter more analytic and information-processing in character (Calfee & Floyd, 1973). Although the literature on teaching effectiveness needs to be approached with caution, one can find consistent signs to support the notion that the analytic-minded teacher is more effective in promoting academic growth, (Potter, 1975; Rosenshine & Furst, 1971). Another instance comes from the work

of Evertson and Brophy (1973): "The teacher who is well organized, who monitors the class regularly and nips potentially serious problems in the bud, and who has well established routines for handling everyday procedural matters tends to be more successful in producing learning gains." This sounds to us like a description of a highly analytic teacher.

Next, we want to highlight three desirable characteristics of tests to be used for short-term instructional decisions:

1. The individual test needs to be "clean," in the sense that demands on the student extraneous to the skill being measured are kept to a minimum. The results from a clean test are much easier to interpret than those from a test where many factors enter in an uncontrolled fashion.
2. Rather than being rigorously standardized, the testing system should permit clinical probing. Such variations in the testing procedure need not be random. We have proposed factorial test designs as a method for systematic exploration of the student's ability to handle a task.
3. Tests for instructional decision-making require more attention to breadth than precision (Cronbach [1970] refers to these as "bandwidth" and "fidelity," respectively). Achieving this goal requires attention to efficiency in the testing procedure, and especially in the choice of where to begin testing for a student.

Each of these issues--clean tests, factorial test design, and efficient entry testing--is a complex matter. We cannot do more below than emphasize a few of the main points.

Clean Tests

A clean test is one in which a single well defined component is examined (Calfee, in press; Calfee, Chapman, & Venezky, 1972). The test begins as simply as possible; ideally, no student should make a mistake under the simplest conditions. This shows the student understands the nature of the test and can handle the general test-taking requirements. Then the difficulty of the test is increased systematically. As errors occur they indicate the nature of the student's problem. Developing a clean test often requires working backwards, asking the question, "What must the student know to be able to succeed in this task?" In answer to the question, "What does a failure mean?" the teacher must make a guess. Based on the guess, the teacher decides how to simplify the test. If the guess was correct and the student is now successful, his problem has been isolated. If he still makes mistakes, the guessing-testing process is pursued further.

The major barriers to a clean test are often the general test requirements. To do well on a test, the student must understand what is expected of him, and must feel encouraged and motivated to do well. Listening carefully and following instructions are important for success, and some students are better at these general skills than are others. Individual or small-group testing makes it easier for the teacher to assure that all students know what they are to do, and makes it more likely that performance will reflect specific rather than general skills. The clinical tester receives the training needed to gain understanding; the classroom teacher may not have had any such training, but he can be aided by guidelines for determining readiness for a test, and suggestions about how to promote readiness.

Factorial Tests

Complementing the notion of a clean test is the idea of factorial test structure. The clean test approach aims toward constancy in all dimensions of the test except one; the factorial approach aims toward systematic variation in several dimensions of the test. Because the concept is new, we will illustrate in Figure 3 how a factorial structure provides a framework for the instructor to think about in testing reading comprehension. One dimension is the nature of the task; oral reading, silent reading with no time pressure, and silent reading with time pressure. As a student becomes competent he should be able to perform well and equally so under all these conditions. A second dimension is the "question mode." How shall the teacher request information from the student after he has finished reading? Perhaps the simplest approach is to ask him direct, literal questions--these can be quite specific or may allow for a more general response to the passage. A recognition test is slightly more difficult, because the student has to read the question and the alternatives, but at least the answers are provided to him. Production and essay tasks demand even more from the student. To summarize a story requires some sophistication, and failure can be traced to any of several possibilities. If performance has been measured under simpler conditions, most of these possibilities can be evaluated. Variation in materials is the third major dimension. It makes quite a difference whether the student is reading a familiar or an unfamiliar topic; difficulty level of vocabulary also makes a difference.

Insert Figure 3 About Here

Envision each student's performance in the multi-dimensional space of Figure 3. The task of the instructor is to locate the student in this space, in the sense that the instructor knows whether the student can perform accurately and quickly in each cell. In fact, one might conceive of testing that aims to trace through the three-dimensional space a line that represents the

boundary between where the student can perform adequately and where he has trouble. Lord's (1974) discussion of "tailored" testing provides a rationale for the unidimensional situation; the multidimensional case remains to be developed, to the best of our knowledge.

Entry Level Assessment

We agree fully with Guszak's (1972) characterization of the good diagnostic reading teacher as someone ". . . capable of making a sequence of relatively simple determinations of a pupil's reading achievement level, his achievement potential, and his prominent skills needs" (p. 22). For the teacher to accomplish this task with any precision, especially when the individual differences within the class are substantial, the teacher must make quick and accurate determinations of the student's level of performance. Starting an assessment in the right "neighborhood" is essential if time is to be used wisely.

Where the teacher has continuing day-to-day knowledge of the student, choosing the proper "entry point" for assessment may be fairly easy. But what about the new student? The new subject matter? The first day of class?

Developing instruments to meet this need seems to us an interesting challenge, and so we will report our experiences--we have little evidence on the reliability of these procedures, though they spring from a well established statistical framework (Wald, 1947).

Here is a systematic but flexible technique for rapidly classifying students whose level of decoding, vocabulary, and comprehension is unknown and may range anywhere from first to eighth grade (Calfee & Hoover, 1974). Choose a few lists of words arranged by difficulty level, and say to the student "Here are some word lists I would like you to read." Which list, A, B, C, D, or E, do you think you can read?" As soon as the student has pointed to the list he thinks he can read, the teacher has a piece of useful information. If

the student's self-assessment agrees with his subsequent performance, he knows realistically what he can do. If he performs two or three levels below his estimate, he at least has a good self-concept.

The teacher then asks the student to read the list he has just pointed to. If he has trouble with several words, the teacher asks him to try an easier list. If he pronounces every work quickly and correctly, the teacher asks him to read a harder list. The student will reach the limit of his skill within a few minutes. A similar procedure is used to assess the level of understanding of word meanings and of paragraph comprehension.

We have used a test built around this model for research activities, and are pleased with the rich return from what is generally less than a twenty-minute test session. But the point to be stressed here is the value of this test for purposes of determining entry level to other tests (and to instruction of course). Precise assessment of a student's skills and knowledge, if it is to be also efficient and not time consuming, requires a quick screening to determine relative standing in different component areas of reading.

Categories of Reading Skills

Reading includes several areas of knowledge and skills and any analytic effort to assess reading must attempt a "first cut" of the collection into reasonably digestible pieces. We have suggested elsewhere (Calfee, Drum, & Arnold, in press) this list: decoding, vocabulary, grammar, transliteral comprehension, and inferential comprehension.

Decoding is the translation from print to sound. It is not clear at what point during the acquisition of reading that the student can best develop this skill. Neither is it clear how decoding skills serve the advanced reader. But a good deal of data exists to support the proposition that the reader of English who can't look at new sets of words and decode them with fluency is likely to have trouble acquiring mastery of other reading skills:

1. Not all reading programs do a good job of training students to decode. Certain approaches are noticeably less effective in promoting the acquisition of decoding skills (Barr, 1974; Chall, 1967).
2. Not all students learn decoding skills in the elementary grades. At the end of the fifth grade many children still evidence lack of skill in handling basic decoding skills (McDonald & Elias, 1975).
3. Substantial correlations are found between decoding skill and school performance up through college (Venezky, 1974b).[4]

The student also needs to be able to define words, to appreciate synonyms, and to recognize common usage of a word. The science question in Figure 4 requires some understanding of the word orifice. The dictionary definition is a start. But few words have a single meaning, and common words have many meanings. Furthermore, even if the student were to internalize the dictionary, society and individuals keep devising idiosyncratic meanings.

Insert Figure 4 About Here

Reading teachers realize that vocabulary development is vitally important to success on academic tasks. Austin and Morrison (1963) reported that more than 75 percent of the teachers in their sample spent "considerable" or "moderate" time in vocabulary development. Rubin, Trismen, Wilder, and Yates (1973) report comparable findings in their survey of teachers in compensatory reading programs. Unfortunately, it is far from clear that the instructional emphasis is accompanied by adequate assessment, sufficient to show not only whether the student "knows," a word, but at what level, and with what degree of fluency.

Some may find it quaint to include grammar as part of the reading process, but it probably has as much place as comprehension skills. In both instances, understanding requires the transfer of skills from oral language to a new context, and the expansion and elaboration of those skills to meet the peculiar demands of the written language (Olson, 1975). An important distinction also

exists between style and substance. Style refers to following the proper convention: producing all the past and plural markers, using proper word order, and the like. If the student is going to speak or write English "properly," he has to know the conventions and use them in the proper context. There are also substantive matters in grammar. Sometimes meaning is disambiguated only when the plural marker, the past marker, or some other morphological ending is noted. If a particular word order has one meaning and a different word order conveys a different meaning, a substantive difference in grammar is apparent. "Bill told Jane to snitch the ice cream" has a different meaning from "Bill was told by Jane to snitch the ice cream." The answer to "Who will be punished for snitching the ice cream?" depends upon recognizing this difference. Many children come to school with adequate knowledge of English syntax; others may need some help. It is the task of instructional assessment to distinguish one group from the other.

Comprehension is a complicated matter; it can be virtually synonymous with thinking. Trying to analyze the process of comprehension is an interesting challenge. We propose here two broad categories of comprehension tasks, transliteral and inferential. Transliteral comprehension requires the student to have meanings for the words, recognition of word order, and either direct or analogical experience with the content, so he can extract and remember information conveyed directly by the passage, information fairly close to the surface. Some questions can be answered by using matching techniques, some by prior experience without reading the passage, and some require an understanding of key terms. Useful assessment procedures sort out the strategies used by students to answer various types of questions.

There is a kind of comprehension that requries a broader and deeper analysis of the textual information. For instance, consider this "comprehension" question:

"Most of the women in the United States are _____.

(a) plumbers, (b) citizens, (c) redheads, or (d) waitresses."

With no passage to read, how does the student select the right answer? The task is only modestly related to reading, though it comes from an actual comprehension test. The student unfamiliar with our culture might think that "redheads" was right; "waitresses" makes sense if many of the women in his experience have been waitresses. An advocate of the women's liberation movement might choose "plumbers." The "correct" answer to the question actually seems stilted and perhaps absurd. The student must rely on knowledge and experience that goes beyond the question and looks at the demand of the task. The good reader brings to bear on the topic what he knows, what he learns from the passage, and what he can figure out about the tester's reasoning and intentions. The teacher needs to know which of these is behind the "poor" student's failure.

The teacher who wishes to "measure comprehension" should be prepared to cover the full range of the student's skills--these include not only finding facts and making simple inferences, but also solving the problem of when to do one or the other. Moreover, the making of inferences is not only a logical process. Many comprehension questions require a process of inference that is more analogical than logical. This requirement seems altogether reasonable, because life experiences are often based more on metaphor than logic. We make comparison with experience and fill in the missing parts of an event by analogy rather than by Aristotelian inference.

The reason for the separation of reading into components like those listed above is straightforward--methods of assessment and selection of instructional

treatment are distinctive for each component. If such is not the case, then the division into components is a useless exercise. The methodology for evaluating the hypothesis that these are independent components--and such a hypothesis is inherent in the listing of the components, we believe--is also straightforward (Calfee & Elman, in press), though only a smattering of research exists currently. We realize that our "shopping list" is not the same as what others might propose; indeed, with more thought and evidence we might want to change it. But we see little point to continued argument about the "fundamental components" in skilled reading and the acquisition of reading. Let researchers move on to propose the systematic, comprehensive, and generalizable research designs necessary to decide which of the many process models are viable. Such research will have theoretical and practical payoff. In the meantime, we might put a moratorium on models with more than 7 ± 2 information-processing stages; these tend to overload the capacity of the reader to understand the model.

Task Requirements in Assessment

In examining these categories of reading skills, we also need to analyze the task requirements for successful performance on a particular test within a given category. Some task requirements are specific to a given area, but others cut across all areas. For instance, the same basic situation may be presented to the reader so that he must recognize the correct answer from a set of alternatives, or must produce the correct answer from memory. The person's skill may allow him to perform well on one form of the task and not on the other. As Kintsch (1970, Ch. 5) notes, different performances under the two task formats permit the researcher (or tester) to infer underlying processes. Recognition of previously studied information suggests the information has been stored adequately; recall suggests that it was stored in a retrievable format.

To find what a person "really" knows, the teacher must devise various ways to tap that knowledge. As noted earlier, it is relatively easy to show that the student cannot remember it under certain conditions. The most direct way to assess a person's knowledge is to ask him a direct question. If he does not give the answer, then a second, more probing question can follow. "Do you think it's this?" Maybe the probe will trip the memory key so the student responds with the correct answer.

Speed and accuracy comprise another important task dimension. Speed is not always "good," but often it is. Automaticity in basic skills can be especially critical (LaBerge & Samuels, 1974). For example, a few years ago we worked with some researchers who were developing a reading series for kindergartners. They had devised an algorithm for teaching children to decode. First the student learned a few letter-sound correspondences, then he moved his finger from one letter to another to blend the sounds: "b;" "b-a, ba;" "ba-t, bat." Within a short time the kindergartners could decode a fairly substantial set of words. Some students were much faster than others, or course. Some could look at the word and say "bat" and others were still going "b-a-t, bat." Then they were asked to read sentences for the first time. The task changed from decoding one word at a time at a relatively easy pace to decoding a whole string of words. Furthermore, the children were expected to answer questions when they finished the sentence. A few seemed to become "instantly dyslexic" at this juncture in the program. In our opinion, this resulted from differences in speed of decoding. Speed of reading single words was not important per se. But it took so long for some students to translate the sentence word-by-word, that by the time they reached the end of the sentence they had forgotten the beginning. Since the decoding strategy didn't work, these students began to guess from initial letters, or they looked at the pictures, searching for meaning with little regard for the print--strategies typical of poor readers.

What is the import of the speed-accuracy distinction for the classroom teacher? Formerly, teachers were encouraged to test for both speed and power. Today, in the era of behavioral objectives and mastery learning, the distinction is largely overlooked. The student who is correct on 80 percent of the items on a multiple-choice test has "mastered" the objective, without regard to how quickly and easily he performs the task, and without regard to how he might perform under different conditions and different demands (e.g., Block, 1974). If the objective is fundamental to the learning of another task, the student may come to grief unless he is fluent with the first objective. In this connection, some evidence has been cited in support of the relative independence of speed of reading and accuracy of comprehension (Gates, 1921; Singer, 1970). Unfortunately, our reading of the evidence leaves us far from convinced about the actual degree of separability of these two measures.

Another point can be mentioned only in passing. Assessment is often most meaningful when carried out in a training context (Calfee, et al., 1971). Short-term training may serve to clarify the task demands for the student. The teacher can note questions and comments by the students as they perform the task. In the State of California, at least one major assessment project includes a pre-test which the teacher is encouraged to give to students until they are thoroughly familiar with how to take the test. Certain commercial tests (e.g., Stanford Achievement Test Battery) also include short practice tests to familiarize the students with the format and type of content they can expect to encounter. This seems a most sensible practice. More generally, the teacher's assessment should aim to measure the student's response to the ongoing instructional program.

Assessment of Transfer

Educators must aim to teach for transfer. Teaching students everything they need to know is impossible. Acquiring knowledge that is transferable generally requires that the student understand principles as well as basic facts. Transfer sometimes happens automatically, but it is often advisable to teach the principle, and then to check or assess whether the principle has actually been acquired. Giving many examples of a principle allows students to have experience with a variety of instances where the principle applies. This procedure means that the teacher must be continually checking not only what students have learned, but also whether the student has attained the principles.

How does one assess the extent of transfer? By changing certain features of the situation from those that existed during training, and seeing whether performance remains stable. By choosing novel instances of a general principle not part of training, and seeing whether the student can apply the principle. By asking the student to state the principle and to supply novel instances exemplifying the principle.

Silberman (1967) demonstrated some years ago the importance of assessment of transfer in the evaluation of a beginning reading program. Teaching students to read a list of words by rote is fairly easy—it may be dull for the teacher and student, but it can be done. However, when Silberman tested for transfer using a variation on the Esper paradigm, in which one portion of a set of associations are learned and transfer is measured by testing other portions of the system (Figure 5), he found that the students had learned what they were taught, nothing more. Using the transfer measure as the standard for a good training program, Silberman proceeded to modify the training program until it worked--until the students learned not only what they were taught, but the principles that allowed them to apply the knowledge in new situations.

Insert Figure 5 About Here

Silberman tested transfer through the Esper paradigm. This is only one of several paradigms developed by experimental psychologists to measure "what is learned" in a deeper sense than simple rote associations (Calfee, 1975b, pp. 393-398; pp. 423-429; Calfee, 1975c; Martinson, in preparation). The advantage of these paradigms is that they provide precise information about what elements of original learning have and have not transferred to a new situation. This precision is in contrast to the vague measures that are all too often used as an index of "transfer" in reading research--the criterion measure is performance on the California Achievement Test, and the transfer measure is performance on the Metropolitan Achievement Test. Whether one observes transfer or not, the exact meaning of the results is uncertain.

Summary

What can the researcher do to help the reading teacher with the task of classroom assessment? In our opinion, this is an area of need that has scarcely been touched. To be sure, many of the new movements in testing seem to have the goal of improving classroom assessment. But the new tests seem quite like the old in appearance and application. The teacher is told not to measure the student's performance against the norms of grade level equivalent or percentile rank. Rather, the teacher should use a criterion--the student must pass 80 percent of the items on a multiple-choice test. But are the items really appropriate? What is the relevant domain for generalization? To what degree does the multiple-choice task relate to other tasks? Why 80 percent--why not 50 percent or 100 percent? How reliable are the data for a particular decision? How valid is the decision?

These are not esoteric questions. They are at the core of the issue of whether it is worth the teacher's and student's time and effort to carry out the assessment.

Conventional "norm-referenced" tests build upon a substantial and well-developed theoretical base. With suitable modification, the same principles can serve in the development of tests for in-class use. The empirical procedures for certifying the adequacy of conventional tests is also well established. Little more is needed for certifying in-class tests, save for the linking of these tests to the instructional base. The norm-referenced test is curriculum-free. The in-class test has to prove its usefulness for making effective and efficient instructional decisions, and for assessing the direct and indirect results of instruction flowing from such decisions.

Carrying out research within this framework will pose special challenges to the behavioral scientist. It requires continuous assessment while the student is engaged in instruction. Computer-assisted instruction solves some problems of control over instruction, and for certain purposes this may be desirable. But most students learn to read in classrooms with a teacher, and it is in this context that we think the greatest payoff will be found. The costs are substantial--the investigator must make himself welcome in the classroom to the point of establishing a collaborative relation with the teacher. The instructional materials and the instructional activities of the teacher need to be monitored and in some instances brought under control. We believe that the payoff can also be considerable: increased knowledge about the cognitive processes that mediate the acquisition of reading skill, and the development of practical assessment tools for more effective teaching of reading.

References

Anastasi, A. Psychological testing. New York: MacMillan, 1968

Atkinson, R. C., & Paulson, J. A. An approach to the psychology of instruction. Psychological Bulletin, 1972, 78, 49-61.

Austin, M. C., & Morrison, C. The first R: The Harvard report on reading in elementary schools. New York: Macmillan, 1963.

Banas, C. New testing method cited for drop in reading scores at public schools. Chicago Tribune, December 4, 1975, pp. 1; 23.

Barr, R. The effect of instruction on pupil reading strategies. Reading Research Quarterly, 1974, 10, 555-582.

Block, J. H. Schools, society and mastery learning. New York: Holt, Rinehart and Winston, 1974.

Bloom, B. S. Stability and change in human characteristics. New York: Wiley, 1964.

Calfee, R. C. The design of experiments and the design of curriculum. Paper presented at Stanford Evaluation Consortium, Stanford University, 1975 (a).

Calfee, R. C. Human Experimental Psychology. New York: Holt, 1975 (b).

Calfee, R. C. Memory and cognitive skills in reading acquisition. In D. Duane & M. Rawson (Eds.), Reading, perception and language. Baltimore, Md.: York Press, 1975(c).

Calfee, R. C. Sources of dependency in cognitive processes. In D. Klahr (Ed.), Cognition and Instruction. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

Calfee, R. C. Assessment of independent reading skills: Basic research and practical applications. In A. S. Reber & D. Scarborough (Eds.), Toward a psychology of reading: The proceedings of CUNY Conference, Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., in press.

Calfee, R. C., Arnold, R. D., & Drum P. A. Book review of E. J. Gibson & H. Levin The psychology of reading. In Proceedings of the National Academy of Education, in press.

Calfee, R. C., Chapman, R. S., & Venezky, R. L. How a child needs to think to learn to read. In. L. W. Gregg (Ed.), Cognition in learning and memory. New York: John Wiley and Sons, 1972.

Calfee, R. C., Cullenbine, R. S., deForcel, A., & Royston, A. B. Further explorations of perceptual and cognitive skills related to reading acquisition. Paper presented to American Psychological Association Convention, 1971

Calfee, R. C., Drum, P. A., & Arnold, R. D. What research can tell the teacher about assessment. In S. J. Samuels (Ed.), What research has to say to the teacher of reading. In press.

Calfee, R. C., & Elman, A. The application of mathematical learning theories in educational settings: Possibilities and limitations. In H. Spada & W. Kempf (Eds.), Structural models of thinking and learning. Bern, Switzerland: Hans Huber, in press.

Calfee, R. C., & Floyd, J. The independence of cognitive processes: Implications for curriculum research. Cognitive processes and science instruction. Bern, Switzerland: Hans Huber, 1973.

Calfee, R. C., & Hoover K. A. RAMOS assessment system. Unpublished mimeographed paper. Stanford University, Stanford, California, 1974.

Calfee, R. C., & Venezky, R. L. Component skills in beginning reading. In K. S. Goodman & J. T. Fleming (Eds.), Psycholinguistics and the teaching of reading. Newark, Delaware: International Reading Association, 1969.

Carver, R. P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518

Chall, J. S. Learning to read: The great debate. New York: McGraw-Hill, 1967.

Cromer, W. The difference model: A new explanation for some reading difficulties. Journal of Educational Psychology, 1970, 61, 471-483.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions (2nd ed.). Chicago: University of Illinois Press, 1965.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, J. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J., Rajaratnam, J., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16(II), 137-163.

Cronbach, L. J., & Snow, R. E. Aptitudes and instructional methods. New York: Irvington, in press.

Educational Testing Service. Cooperative primary reading test. Princeton, N.J.: Educational Testing Service, 1970.

Evertson, C. M., & Brophy, J. E. High-inference behavioral ratings as correlates of teaching effectiveness. Austin, Texas: Research and Development Center for Teacher Education, University of Texas, 1973.

Farr, R., Tuinman, J. J. The dependent variable: Measurement issues in reading research. Reading Research Quarterly, 1972, 7, 413-429.

Feshbach, S., Adelman, H., & Fuller, W. W. Early identification of children with high risk of reading failure. Paper presented to American Educational Research Association, New Orleans, 1973.

Gates, A. I. An experimental and statistical study of reading and reading tests. Journal of Educational Psychology, 1921, 12, 303-314.

Gibson, E. J., & Levin, H. The psychology of reading. Cambridge, Mass.: The MIT Press, 1975.

Green, D. B. The nature and use of criterion-referenced and norm-referenced achievement tests. Special report by the Association of California School Administrators, Vol. 4, No. 3, 1976.

Cuszak, F. J. Diagnostic Reading Instruction in the Elementary School. New York: Harper & Row, 1972.

Hively, W. (Ed.), Domain-referenced testing. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Holland, J. G. Variables in adaptive decisions in individual instruction. Technical Report, Learning Research and Development Center, University of Pittsburgh, 1975.

Kintsch, W. Learning, memory and conceptual processes. New York: Wiley, 1970.

Knapp, T. R. An application of balanced incomplete block designs to the estimation of test.norms. Educational and Psychological Measurements, 1968, 28, 265-272.

LaBerge, D., & Samuels, S. J. Toward a theory of automatic information processing in reading. Cognitive Psychology, 1974, 6, 293-332.

Levine, M. The academic achievement test: Its historical context and social functions. American Psychologist, 1976, 31, 228-238.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology Vol. II: Measurement, psychophysics and neural information processing. San Francisco: W. H. Freeman and Company, 1974

Lorton, P. Jr. Computer-based instruction in spelling: An investigation of optimal strategies for presenting instructional material. Unpublished doctoral dissertation, Stanford University, 1972.

Martinson, L. The acquisition of decoding skills in beginning readers. Unpublished doctoral dissertation, Stanford University, in preparation.

McCullough, C. M. Response of elementary school children to common types of reading comprehension questions. Journal of Educational Research, 1957, 51, 65-70.

McDonald, F. J., & Elias, P. Beginning teacher evaluation study: Phase II final report (vol. I: Ch. 10). Princeton, N. J.: Educational Testing Service, 1975.

Mellon, J. C. National assessment and the teaching of English. Urbana, Ill. National Council of Teachers of English, 1975.

Olson, D. R. A review of Toward a literate society, edited by John B. Carroll & Jeanne Chall. Proceedings of the National Academy of Education, 1975, 2, 109-178.

Perfetti, C. A., & Hogaboam, T. The relationship between single word decoding and reading comprehension skill. Journal of Educational Psychology, 1975, 67(4), 461-469. (a)

Perfetti, C. A., & Hogaboam, T. The effects of words experience on decoding speeds of skilled and unskilled readers. Paper presented at the annual meeting of the Psychonomics Society, Denver, Colorado: November 1975. (b)

Potter, D. A critical review of the literature: Teacher performance and pupil growth. In. F. J. McDonald & P. Elias (Eds.), Beginning teacher evaluation study: Phase II Report. Princeton, N.J.: Educational Testing Service, 1975.

Rosenshine, B., & Furst, N. Research in teacher performance criteria. In B. O. Smith (Ed.), Research in teacher education: A symposium. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Rubin, D., Trismen, D. A., Wilder, C., & Yates, A. A descriptive and analytic study of compensatory reading programs. Phase I Report Contract No. OEC-71-3715, Educational Testing Service, Princeton, New Jersey, 1973.

Silberman, H. F. Experimental analysis of a beginning reading skill. In J. P. DeCecco (Ed.), The psychology of language, thought, and instruction. New York: Holt, Rinehart and Winston, 1967. Also in Programmed Instruction, 1964, 3, 4-8.

Singer, H. Research that should have made a difference. Elementary English, 1970, 47, 27-34.

Sticht, T. G. (Ed.). Reading for working: A functional literacy anthology. Alexandria, Virginia: Human Resources Research Organization, 1975.

Sticht, T. G., Caylor, J. S., Kern, R. P., & Fox, L. C. Project REALISTIC: Determination of adult functional literacy skill levels. Reading Research Quarterly, 1971, 7, 424-465.

Thorndike, R. L. Dilemmas in diagnosis. In W. H. MacGinitie (Ed.), Assessment problems in reading. Newark, Delaware: International Reading Association, 1973.

Venezky, R. L. Testing in reading: Assessment and instructional decision-making. Urbana, Illinois: National Council of Teachers of English, 1974 (a).

Venezky, R. L. Theoretical and experimental bases for teaching reading. In T. Sebeok (Ed.), Current Trends in Linguistics, Vol. 12. The Hague: Mouton & Co., 1974 (b).

Venezky, R. L. The curious role of letter names in reading instruction. Visible Language, 1975, 9, 7-23.

Walker, D. F., & Schaffarzick, J. Comparing curricula. Review of Educational Research, 1974, 74, 83-111.

Wald, A. Sequential analysis. New York: Wiley, 1947.

Weintraub, S., Robinson, H. M., Smith, H. K., & Roser, N. Summary of investigations relating to reading July 1, 1972 to June 30, 1973. Reading Research Quarterly, 1973, 9, 247-513.

Weintraub, S., Robinson, H. M., Smith, H. K., Pleasas, G. S., & Rowls, M. Summary of investigations relating to reading July 1, 1973, to June 30, 1974. Reading Research Quarterly, 1974, 10, 267-543.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

Table 1

Example of Student-Item Matrix

with Consistent Items

(0=correct, 1=error)

		Items				Student
		1	2	3	4	Total Score
Students	A	1	1	1	1	4
	B	0	1	1	1	3
	C	0	1	1	1	3
	D	0	0	0	1	1
	E	0	0	0	1	1
	F	0	0	0	0	0
Item Totals		1	3	3	5	

Table 2

Analysis of Variance of Student-Item Matrix, Estimation of Variance Components, and Calculation of Reliability

1. Analysis of variance summary table

Source	df	MS	EMS
Students	5	.600	$\sigma^2_{SI} + \sigma^2_S$
Items	3	.433	$\sigma^2_{SI} + \sigma^2_I$
SI	15	.113	σ^2_{SI}

2. Estimation of variance components

$$\sigma^2_S = MS(S) - MS(SI) = .487$$

$$\sigma^2_{SI} = MS(SI) = .113$$

3. Reliability of contribution of each item to individual differences in student's total score

$$\alpha = \frac{\sigma^2_S}{\sigma^2_S + \sigma^2_{SI}} \quad \frac{.487}{.487 + .113} = .81$$

Table 3

Example of Student-Item Matrix

with Less Consistent Items than those in Table 1,

Showing Analysis of Variance

and Estimation of Reliability

Student-Item Matrix

	Items				Student
	1	2	3	4	Total Score
A	1	1	1	1	4
B	0	1	1	1	3
C	1	0	1	1	3
D	1	0	0	0	1
E	0	1	0	0	1
F	0	0	0	0	0
Item Totals	3	3	3	3	

Table 4

Student-Item Matrix with Items Grouped

According to Test Factor T

Item	1	2	3	4	5	6	7	8	Student Total Score	Subtest Totals: Items 1 - 4	Subtest Totals: Items 5 - 8
1	1	1	0	1	0	0	0	0	3	3	0
2	0	1	0	0	1	1	1	1	5	1	4
3	1	0	0	0	0	0	0	0	1	1	0
4	1	1	1	1	0	1	1	1	7	4	3
5	1	1	1	1	0	0	0	1	5	4	1
6	1	0	0	0	1	1	1	1	5	1	4
7	1	1	1	1	1	1	1	1	8	4	4
8	1	0	0	0	0	0	0	0	1	1	0
9	0	0	0	0	1	0	1	1	3	0	3
10	1	1	1	1	0	0	1	0	5	4	1
11	1	1	0	1	1	1	1	1	7	3	4
12	1	1	1	1	1	0	1	1	7	4	3
13	0	0	0	0	0	0	0	0	0	0	0
14	0	0	1	0	1	1	1	1	5	1	4
15	1	1	1	0	0	0	0	0	3	3	0
16	1	1	1	1	0	1	0	0	5	4	1
17	1	1	0	1	1	1	1	1	7	3	4
18	0	0	0	0	0	0	0	1	1	0	1
19	0	0	0	0	0	0	1	0	1	0	1
20	0	0	0	0	0	1	1	1	3	0	3
Item Total Score	13	11	8	9	8	9	12	12			

Table 5

Analysis of Variance
of Original Student-Item Matrix,
Estimation of Variance Components
and Computation of Overall Reliability

1. Analysis of Variance

Source	df	MS
Student	19	.749
Item	7	.196
SI	133	.183

2. Estimation of Variance Components

σ^2_S = MS(S) - MS(SI) .566 σ^2_{SI} = .183

3. Reliability of Item Contribution to Total Score

$$\alpha = \frac{.566}{.566 + .183} = .756$$

Table 6

Analysis of Variance,
Estimation of Variance Components,
and Calculation of Reliability Indices
for Total and Subtest Scores

1. Analysis of Variance

Source	df	MS	EMS
Students	19	.749	$\sigma^2_{SI(T)} + \sigma^2_{ST} + \sigma^2_{S}$
Tests	1	.0	$\sigma^2_{SI(T)} + \sigma^2_{ST} + \sigma^2_{I(T)} + \Sigma^2_{T}$
ST	19	.645	$\sigma^2_{SI(T)} + \sigma^2_{ST}$
Items (T)	6	.229	$\sigma^2_{SI(T)} + \sigma^2_{I(T)}$
SI(T)	114	.106	$\sigma^2_{SI(T)}$

2. Reliability of Subtest Contribution to Total Score

$\sigma^2_S = MS(S)-MS(ST) = .104 \quad \sigma^2_{ST} = MS(ST)-MS(SI(T)) = .539$

$$\alpha = \frac{.104}{.104 + .539} = .162$$

3. Reliability of Item-Within-Subtest Contribution to Subtest Scores

$\sigma^2_{ST} = .539 \quad \sigma^2_{SI(T)} = .106$

$$\alpha = \frac{.539}{.539 + .106} = .836$$

Note: σ^2 is a random effect, Σ^2 is a fixed effect in the analysis of variance model.

Figure Captions

Figure 1. Frequency distribution of kindergarten alphabet scores for total sample, for students above, and for students below grade level in reading achievement at end of first grade (Calfee, in press).

Figure 2. Cut-point results based on kindergarten alphabet scores and first grade reading achievement (Calfee, in press).

Figure 3. A factorial structure on dimensions of reading for instructions and assessment.

Figure 4. Sample science test item with illustration.

Figure 5. Illustration of training and transfer matrix used by Silberman (1967) for assessment of decoding principles in beginning reading curriculum.

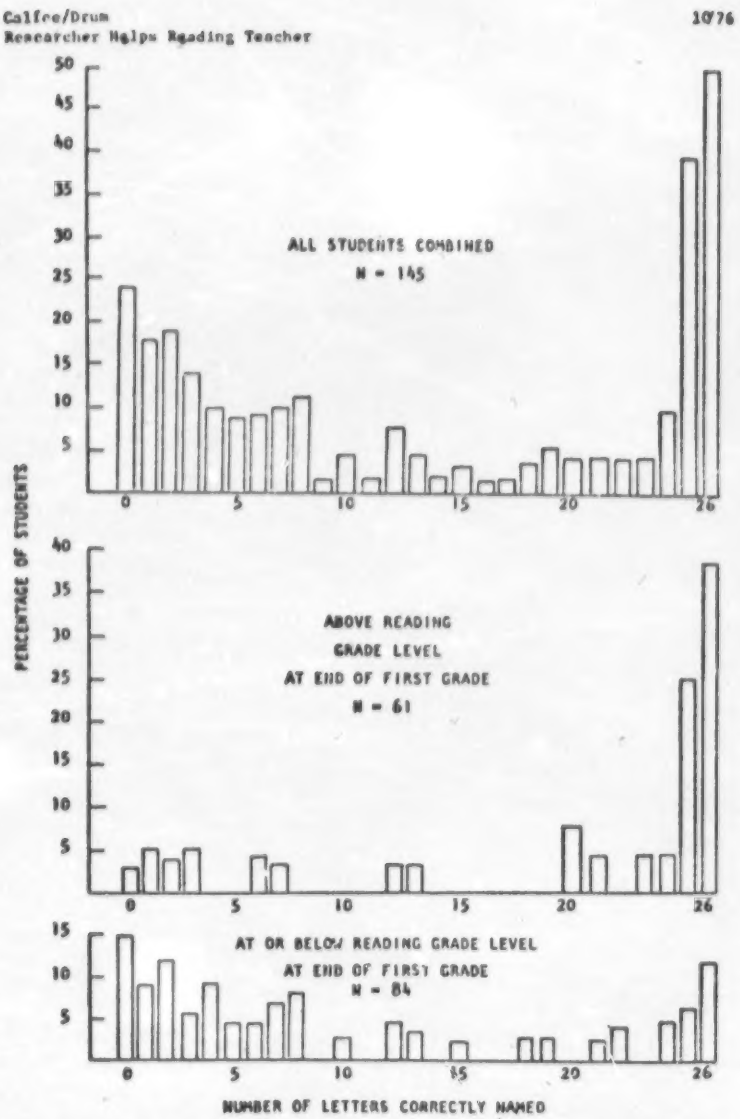


Figure 1. Frequency distribution of kindergarten alphabet scores for total sample, for students above, and for students below grade level in reading achievement at end of first grade (Calfee, in press).

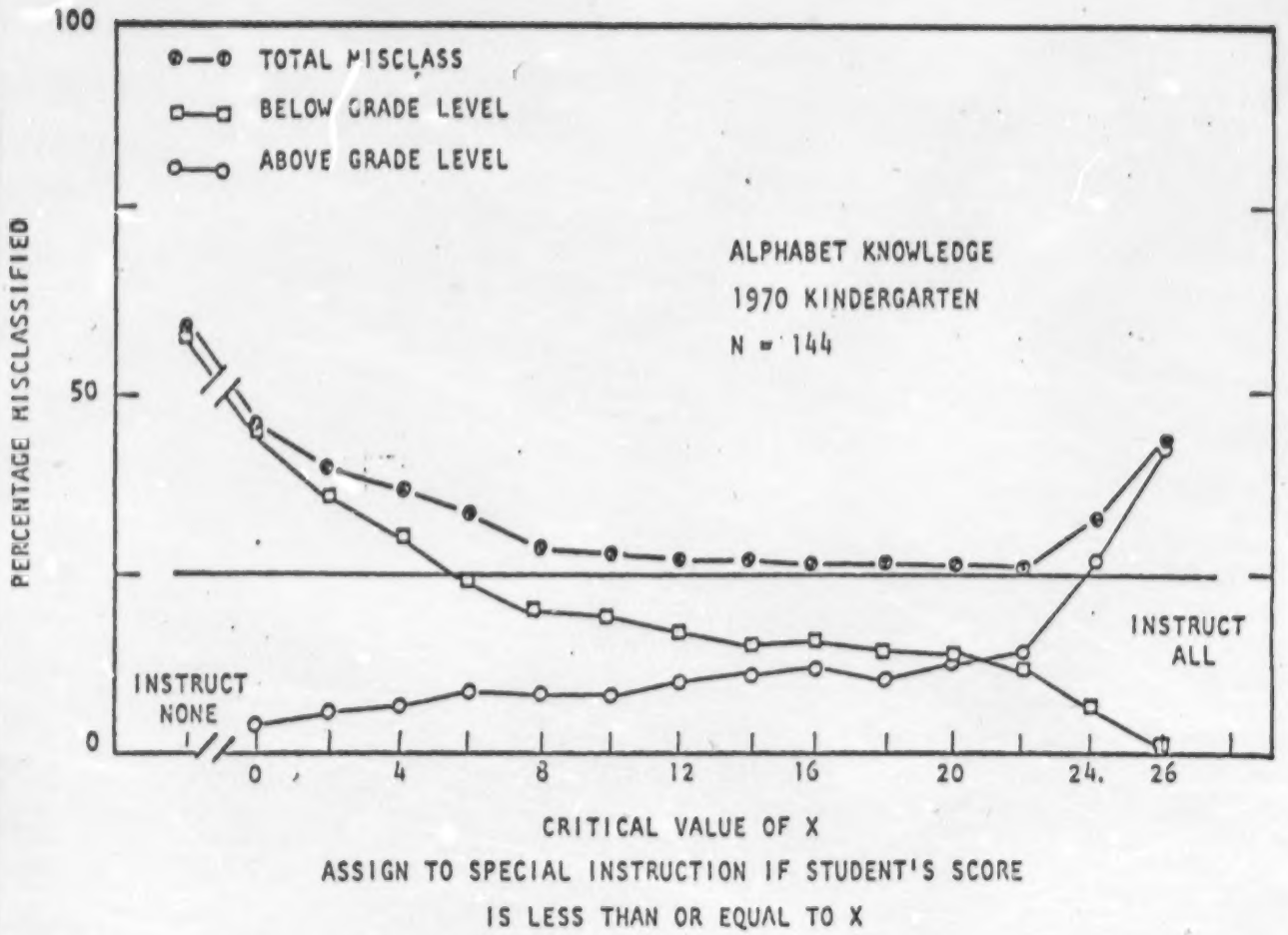


Figure 2. Cut-point results based on kindergarten alphabet scores and first grade reading achievement (Calfee, in press).

TASK		MATERIALS			
		Familiar Topic		Unfamiliar Topic	
Reading Mode	Question Mode	Easy Vocabulary	Difficult Vocabulary	Easy Vocabulary	Difficult Vocabulary
Read Aloud	Oral/ Literal Recognition/ Literal Recognition/ Interpretative Production/ Essay				
Read Silently No Time Pressure	Oral/ Literal Recognition/ Literal Recognition/ Interpretative Production/ Essay				
Read Silently Time Pressure	Oral/ Literal Recognition/ Literal Recognition/ Interpretative Production Essay				

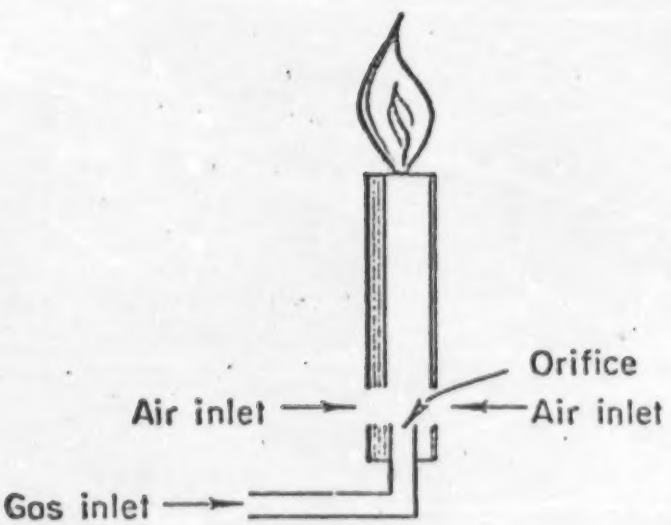
Figure 3. A factorial structure on dimensions of reading for instructions and assessment

VOCABULARY = DEFINING, KNOWING, SYNONYMS, RECOGNIZING USAGE

WHAT IS AN ORIFICE?

"A MOUTH OR SIMILAR OPENING; A HOLE; AN APERTURE"

WHAT IS THE ORIFICE IN THIS PICTURE?



Typical gas burner

Figure 4. Sample science test item with illustration

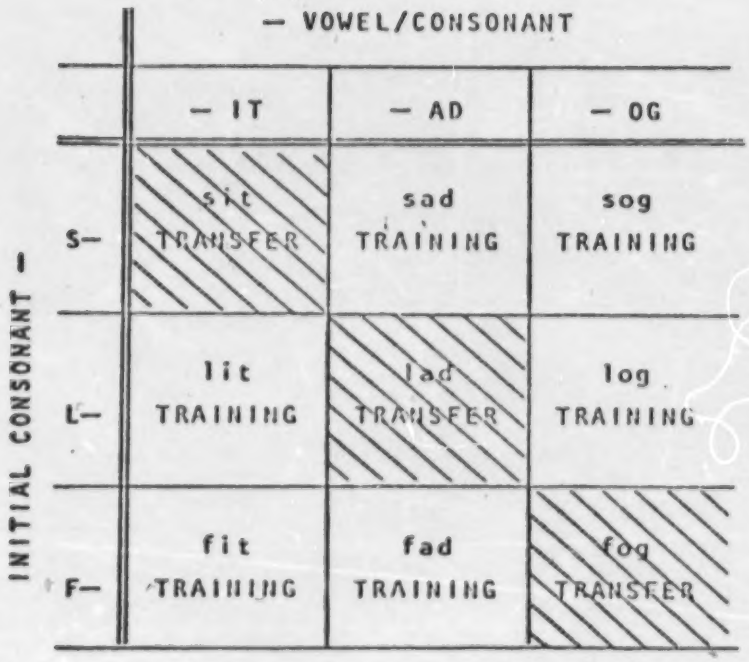


Figure 5. Illustration of training and transfer matrix used by Silberman (1967) for assessment of decoding principles in beginning reading curriculum

Footnotes

[1]Domain-referenced testing probably comes closest in spirit to the conceptualization that seems most useful to us. Theory and practice remain to be established for domain-referenced tests, though some interesting beginnings exist (Hively, 1974, especially chapters by Millman, Miller, and Nitko; Knapp, 1968).

[2]McCullough (1957) has presented evidence for independence of comprehension processes in the form of low to moderate correlations between elementary students' responses to comprehension questions about details, main idea, sequence, and creative reading. Unfortunately, the number of items was small, and test reliabilities were not reported. Thus, the modest size of the correlations is not strong evidence of independence, though the data are suggestive.

[3]Holland (1975) has given thought to desirable characteristics of tests for instructional decision-making, and presents some interesting indices:

(a) What proportion of the instructional time is used by testing versus teaching?

(b) Does the test provide useful information for sorting students into instructional groups; if the test results say "assign all students to instruction A," the test has served no useful role for making a decision.

(c) Does the test promote valid decisions; does the student who passes the test succeed without instruction, and contrariwise?

Holland's methods of analysis are fairly crude, but it seems to us that the questions are right. His conclusions about the usefulness of several instructional systems are generally disappointing, but seem to us based on too little data and too superficial an analysis.

Footnotes (continued)

[4]Laboratory research from several sources demonstrates the important relation between fluent skilled decoding and comprehension (e.g., Perfetti & Hogaboam, 1975(a), 1975(b) ; Cromer, 1970; following the analysis in Calfee, Arnold, & Drum, 1976). To be sure, the training studies needed to establish causality remain to be done. It is far from clear that the teaching of decoding skills in regular classrooms receives the emphasis that some reports suggest. For instance, in questioning teachers whose classrooms included some kind of compensatory reading program, it was found that less than one in five teachers of sixth grade students made any extensive use of phonics curriculum programs (Rubin, Trismen, Wilder, & Yates, 1973). More than 95 percent of the teachers at all grade levels said that comprehension was a major goal. Another piece of information from this study bears on the relative emphasis on decoding skills: In second grade, 75 percent of the teachers report that each child reads aloud to an adult once a week or more often. By fourth grade, only 63 percent of the teachers report this much oral reading, and by sixth grade the figure is 57 percent.

OPEN DISCUSSION OF CALFEE PRESENTATION

E. SMITH: Bob, I missed something. You said that you don't have to worry about the reliability of the individual test, but you do have to worry about the reliability over a set of administration errors. How are you going to get reliability over a set, if you haven't got reliability in any of the elements in the set?

CALFEE: That's a technical question, and one of these days I will write a technical answer to it, but basically the answer is going to take this form: Look at a complex factorial test structure; time can be one of the dimensions, as can production and recognition. Imagine a test, materials that may have 40 or 50 items in it, where you maybe have only two or three or four items in a single cell. A way of measuring a reliability within a cell--which is where you ought to be measuring it--is to compute the mean square residual error after you have extracted all of the systematic variance.

As Cronbach points out in his analysis of reliability, whatever is left over is a measure of the reliability of that test. So indeed the technical knowledge for answering that question exists, but that's not part of what we want to say to the classroom teacher. Another part of that answer goes back to the extremely reliable test, where the pattern is either all successes or all errors. If you design a clean test that is aimed at the specific skill, you are very often going to get performance that looks very much like that, so it becomes, manifestly, within the cell, reliable.

We have developed tests that fall within a cell, that consist of six items. We ask the student to pronounce several words when we are looking for performance of a particular character. We are not asking: "Did you get the word right or wrong?" That requires several skills in itself. Instead, we are asking: "How did you handle the 'ou' in about?" We score just that, right or wrong.

We give the first item to the student, and if he or she makes a mistake, we say, "Gee, did you really understand what we are talking about? Because the right answer to that is this, and this is why." Then we try the second item. If he or she misses that, we stop the testing right there. That's all of the evidence we want that they don't know how to handle that test.

If they get one or the other of those right, we give them four more items. We find that they either are right on three or four of them. A small number get three, maybe 1% or 2% or 5% will get a couple right, a couple wrong, a small number get one, and only one right.

You get mostly a pattern where they get them all right or all wrong. The reliability problem can be treated in the most trivial way.

If you really wanted, you could probably begin to pick me apart on some of the details. Making the system really work is going to take more than the hand waving. But the technical background is available in Cronbach's theory of generalizability. For the experimental psychologists, who are not aware of that, let me say it is a fundamentally important work that is going to change your concepts of reliability greatly over the next decade.

CAZDEN: What is the reference to that?

CALFEE: Cronbach, Gleser, Nanda, and Rajaratnam, The dependability of behavioral measurements: Theory of generalizability.

BLOCK: How do you see the interface between the outcomes of testing and what the teacher can do instructionally for a given child? It's very nice to have tailored assessment devices, but if we can't differentiate descriptions as a function of those decisions, what good are they? How do you see that interface working out?

CALFEE: When I think about it seriously, I say that first we have to divide what we mean by reading into a small number of coherent areas. Probably the research ought to aim at one of those at a time, and if it were up to me, I would try to answer that question for decoding. I believe you can work on the answer to the question that you have asked by looking at decoding as a separate problem, independent of the other areas of reading. If we do get a model for answering that question for this one area, we would be in much better shape to know how to solve it for comprehension and for vocabulary development.

If I am wrong in the way I am carving up reading, or if I am wrong in the basic assumption that reading is a bunch of separable skills, the research isn't going to turn up anything very interesting. But what I would want to do is say, "Okay, take decoding, let's carve decoding up into a small number of coherent areas. Let's ask, what are the major dimensions of curriculum development?" They are what will become the segments of the curriculum. If you are really interested, I will send you a paper.

BLOCK: I really am, because I always find it difficult.

CALFEE: What is the means of delivery, because you can teach the same thing in many ways. What are the factors, and what are some reasonable levels of those factors? There is a lot of thinking about the curriculum before you do anything. Now, let's think about teachers. The teacher is not a homogeneous entity, far from it. Teachers come in a variety of forms, and I would hate to do research on teachers any more without including teacher training as part of the design. So dimensions of teachers and dimensions of teacher training programs are important.

Then I would say, "Let me begin a design process. Let me try to get a design that might use 20 or 30 teachers, over the course of a year, in a fairly well controlled, but natural, situation. And let's collect data consistently." What I am talking about is do-able. You have to have good political relations with teachers and teacher units. I collect the data, and I look at it for a year, and then I know how to do the next experiment. The outcome would be a validation of certain training programs for teachers, appropriate for certain kinds of classrooms and students, with answers about where the important curriculum decisions have to be made. We are not going to come up with a curriculum, but you are going to know how to use the chunks of curriculum you have.

What Cronbach points out is something that Herb Clark has also pointed out: There are fixed effects in this business, but there are also random effects, and you can control both of those.

GORDON: I found myself in enthusiastic agreement with most of the points you were making, but when you came to your summary, you, at least by implication, introduced a contradiction. I think you suggest that standardized testing is a terribly useful and dependable device, so that you didn't have much argument with

it. I think that's true. But if we take seriously the things that you are talking about and begin to achieve these, we are going to change the conditions of learning. That introduction to your summary ought to indicate that under traditional conditions, under unchanged conditions, or in the absence of success at the things you are talking about, that prediction holds. If you succeed in the things are are talking about, we are going to change the validity of those predictions.

CALFEE: That's right. It is not a conclusion that I am unaware of. Let me again refer to my own teaching. The poor students in my classes are getting tested every week. I have tutors, who are assigned to help people, and there are fixed standards, so I have a good standardized testing procedure. There are exams, big exams, that ask for mastery of statistical concepts in a global way. The student has got to "get it all together." The standards are fixed, unlike standardized tests. They are a rat race, a treadmill that gets faster as the norms go up. I don't use that approach, and I think that's the answer to the contraduction that you refered to. Following the national assessment model, if everybody did perfectly on what seemed to us to be a reasonable set of general items, then who cares about norms any more? That would radicalize the testing business.

RESNICK: And the education business.

CALFEE: And the education business. I face that in my class by setting absolute standards. If you get 90% of all of the points on all of the exams, you get an A. And they are tough exams. I can readily get evidence on that from the students. Something like 85% of the students in my courses get an A, and it's

not because I am grading easy. They learn.

RESNICK: That general issue is a good one to close on; it leads to some radical and hopeful thoughts for the future.

END SESSION